

1990.6.12  
(312) 360 0080

JCS976 U.S. PTO  
09/1761221  
01/16/01

日本国特許庁  
PATENT OFFICE  
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日  
Date of Application:

2000年5月26日

出願番号  
Application Number:

特願2000-155867

出願人  
Applicant(s):

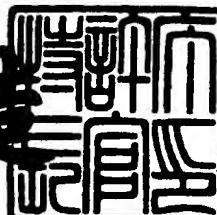
富士通株式会社

CERTIFIED COPY OF  
PRIORITY DOCUMENT

2000年10月27日

特許庁長官  
Commissioner,  
Patent Office

及川耕造



出証番号 出証特2000-3089277

【書類名】 特許願  
【整理番号】 0050852  
【提出日】 平成12年 5月26日  
【あて先】 特許庁長官殿  
【国際特許分類】 G06F 17/30  
【発明の名称】 文書情報検索装置、方法及び文書情報検索プログラムを  
格納したコンピュータ可読の記録媒体  
【請求項の数】 10  
【発明者】  
【住所又は居所】 神奈川県川崎市中原区上小田中4丁目1番1号 富士  
通株式会社内  
【氏名】 阿部 静一郎  
【特許出願人】  
【識別番号】 000005223  
【氏名又は名称】 富士通株式会社  
【代理人】  
【識別番号】 100079359  
【住所又は居所】 東京都港区西新橋3丁目25番47号 清水ビル8階  
【弁理士】  
【氏名又は名称】 竹内 進  
【電話番号】 03(3432)1007  
【選任した代理人】  
【識別番号】 100093584  
【住所又は居所】 東京都港区西新橋3丁目25番47号 清水ビル8  
階  
【弁理士】  
【氏名又は名称】 宮内 佐一郎  
【電話番号】 03(3432)1007

【手数料の表示】

【予納台帳番号】 009287

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9704823

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書情報検索装置、方法及び文書情報検索プログラム  
を格納したコンピュータ可読の記録媒体

【特許請求の範囲】

【請求項1】

ネットワークを経由した検索要求に基づいて文書情報を検索して応答する文書情報検索装置に於いて、

前記検索要求元に、検索条件としてファイルを指定し、指定したファイル内容をネットワークを経由して送信する検索条件指定部を設け、

検索側に、前記検索条件指定部から送信されたファイル内容からキーワードを生成してデータベースから類似文書を検索する検索マシンを設けたことを特徴とする文書情報検索装置。

【請求項2】

請求項1記載の文書情報検索装置に於いて、

前記データベースは、検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存し、

前記検索マシンは、

検索要求に伴って受信したファイル内容からテキスト文を抽出するテキスト抽出処理部と、

前記テキスト文の形態素解析により名詞を抽出する形態素回析部と、

前記名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部と、

前記キーワードによる検索データベースの検索で類似する文書を検索してクラスタントに検索結果を通知する検索実行部と、  
を備えたことを特徴とする文書情報検索装置。

**【請求項3】**

請求項2記載の文書情報検索装置に於いて、前記キーワード生成部は、各名詞が前記文書データベースに格納した検索文書毎のインデックス中の何文書に出現するかの出現数をカウントし、所定範囲の出現数をもつ上位の所定数の単語を選択してキーワードを生成することを特徴とする文書情報検索装置。

**【請求項4】**

請求項3記載の文書情報検索装置に於いて、前記キーワード生成部は、インデックス中の文書数Nとした場合、出現数Hが

$$2N/3 \geq H \geq 1$$

の範囲の出現数をもつ上位の10個の単語を選択してキーワードを生成することを特徴とする文書情報検索装置。

**【請求項5】**

請求項1記載の文書情報検索装置に於いて、前記キーワード生成部は検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索させることを特徴とする文書情報検索装置。

**【請求項6】**

検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存しているデータベースと、

前記検索データベースに登録されていない文書ファイルを検索条件に指定したネットワークからの検索要求によって受信したファイル内容からテキスト文を抽出するテキスト抽出処理部と、

前記テキスト文の形態素解析により名詞を抽出する形態素回析部と、

前記名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部と、

前記キーワードによるデータベースの検索で類似する文書を検索して要求元に検索結果を通知する検索実行部と、

を備えたことを特徴とする文書情報検索装置。

【請求項7】

請求項6記載の文書情報検索装置に於いて、前記キーワード生成部は検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索することを特徴とする文書情報検索装置。

【請求項8】

ネットワークを経由した検索要求に基づいて文書情報を検索して応答する文書情報検索方法に於いて、

検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎にデータベースに保存し、

検索要求元で検索条件にファイルを指定した場合に、指定したファイル内容を検索要求と共にネットワークを経由して検索先に送信し、

検索側で、検索要求に伴って受信したファイル内容からテキスト文を抽出すると共にテキスト文の形態素解析により名詞を抽出し、次に名詞の中から重要語を抽出して論理和でつなげたキーワードを生成し、該キーワードによるデータベースの検索で類似する文書を検索して検索結果を応答することを特徴とする文書情報検索方法。

【請求項9】

請求項8記載の文書情報検索方法に於いて、検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索することを特徴とする文書情報検索方法。

【請求項10】

文書ファイルを検索条件に指定した検索要求を受信するステップと、

検索要求に伴って受信したファイル内容からテキスト文を抽出するステップと

テキスト文の形態素解析により名詞を抽出するステップと、  
名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するステッ  
プと、  
前記キーワードによるデータベースの検索で類似する文書を検索して要求元に  
検索結果を通知するステップと、  
を備えた検索プログラムを格納したコンピュータ可読の記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、大量の文書データの中から必要な文書を迅速に探し出すための文書情報検索装置、方法及び文書情報検索プログラムを格納したコンピュータ可読の記録媒体に関し、特に、文書ファイルそのものを検索条件に指定するという簡単な操作で内容が類似する文書を探し出す文書情報検索装置、方法及び文書情報検索プログラムを格納したコンピュータ可読の記録媒体に関する。

【0002】

【従来の技術】

従来、ネットワーク環境を利用した文書管理システムにあっては、インターネットやイーサネット上に存在する大量の文書データから必要な文書を検索してすばやく参照することのできる文書情報検索装置を提供している。

【0003】

この場合の文書検索は、ユーザが必要とする文書に含まれていると思われる1又は複数の適当な単語や文字列をキーワードとして指定し、この指定したキーワードの単語を含む文書を検索データベースから検索し、文書一覧を検索結果として表示する。

【0004】

この文書情報検索装置にあっては、ネットワーク上に存在する検索対象文書について、その内容から重要語を抽出して列挙したインデックスを文書毎に作成し

て検索データベースに保存している。そしてユーザからキーワードを指定した検索要求があれば、検索データベースのインデックスを検索して文書一覧の検索結果を出すようにしている。

#### 【0005】

更に、従来の文書情報検索装置は、ユーザがキーワード指定で検索した文書一覧の中から必要と思われる文書を検索した後、文書一覧の中から選んだ文書について類似文書検索を指定すると、検索文書の中に出現する頻度の高い用語が自動的に抽出され、前回実行された検索条件に論理和の条件で付加され、類似文書の検索を行うことができる。

#### 【0006】

##### 【発明が解決しようとする課題】

しかしながら、ユーザが電子メールやインターネットで、興味ある文書を入手し、この文書に類似した内容の文書を検索したい場合、現状では入手した文書に含まれている単語や文字列を選んでキーワードにいちいち指定し、まず検索結果として文書一覧を得る。次に、検索した文書一覧の中から文書を選択して類似文書検索を指定して類似文書の検索を行わなければならない。

#### 【0007】

即ち、電子メールやインターネットで入手した文書の類似検索を行おうとしても、従来の文書情報検索装置は、既に検索データベースに登録されている文書しか、文書を検索条件に指定した類似文書の検索はできず、ユーザが電子メールやインターネットで入手した文書を検索条件に使って直接的に類似文書の検索を行うことができない。

#### 【0008】

このためユーザが電子メールやインターネットで入手した文書の中から、文書検索に必要と思われるキーワードを選んで検索条件として入力する必要があり、キーワードが多くある場合は入力に手間がかかる。またキーワードの指定が十分でないと検索漏れを生じ、期待した検索結果が得られない場合がある。

#### 【0009】

更に文書一覧として得られる検索数が膨大となることもあり、文書一覧から関連すると思われる文書を開いて必要な文書を捜し出す大変な手間がかかる場合がある。

#### 【0010】

本発明は、検索データベースに登録されていない文書に類似した文書の検索を簡単な操作ですばやくできる文書情報検索装置、方法及び文書情報検索プログラムを格納したコンピュータ可読の記録媒体を提供することを目的とする。

#### 【0011】

##### 【課題を解決するための手段】

図1は本発明の原理説明図である。本発明は、クライアント12等からのネットワークを経由した検索要求に基づいてサーバ10等の検索側で文書情報を検索して応答する文書情報検索装置であって、クライアント12等の要求元に、検索条件にファイルを指定した場合に、指定したファイル内容をネットワークを経由して送信する検索条件指定部26を設け、サーバ10等の検索側に、検索条件指定部26から送信されたファイル内容からキーワードを生成して類似文書を検索する検索マシン20を設けたことを特徴とする。

#### 【0012】

このため電子メールやインターネット等で興味のある内容を含む文書を入手し、この文書に類似した内容の文書を検索したい場合等に、文書の指定によりアップロードされたファイルを検索条件に指定することで、内容が類似する文書を検索することができる。このためデータベース登録されていない文書であっても自由に検索条件として指定することができ、手間のかかる文書内容に基づいたキーワードの入力を不要とし、簡単且つ迅速に類似文書を探し出すことができる。

#### 【0013】

検索要求元の検索条件指定部26は、指定されたファイル内容の先頭ファイル部分を送信する。通常、文書検索に必要な重要なキーワードは文書の先頭部分に多く存在することから、ファイル内容の先頭部分だけ、例えば先頭の1KB部分を検索条件として送信する。また検索条件に使用する文書ファイルのサイズは様

々であることから、検索条件として送信するファイル容量を決めることで、通信負荷と検索側の処理を軽減する。

#### 【0014】

検索条件指定部26は検索条件として指定するファイルにHTMLファイル及びエクセルファイルを含む。勿論、これ以外のファイル形式であっても、テキスト文書の抽出が可能なファイルであれば、任意のファイル形式のものを含む。

#### 【0015】

サーバ10側の検索マシン20には、検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存したデータベース22が設けられる。また検索マシン20のファイル指定検索部30は、検索要求に伴って受信したファイル内容からテキスト文を抽出するテキスト抽出処理部36、テキスト文の形態素解析により名詞を抽出する形態素回析部38、名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部40、及びキーワードによる検索データベース22の検索で類似する文書を検索してクライアントに検索結果を通知する検索実行部42を備える。

#### 【0016】

キーワード生成部40は、各名詞が検索データベース22に格納した検索文書毎のインデックス中の何文書に出現するかの出現数Hをカウントし、所定範囲の出現数Hをもつ上位の所定数の単語を選択してキーワードを生成する。

#### 【0017】

キーワード生成部40は、インデックス中の文書数Nとした場合、例えば出現数Hが

$$2N/3 \geq H \geq 1$$

の範囲の出現数をもつ上位の10個の単語を選択してキーワードを生成する。これによりデータベースのインデックスに登録している既存文書の類似検索に必要な重要語を絞り込み、類似検索の精度を高める。

#### 【0018】

更にキーワード生成部40は、検索要求に伴って受信したファイルから抽出したプロパティ情報をキーワードに含めて検索させる。この場合のプロパティ情報

は、検索要求に伴って受信したファイルの作成者、文書タイトル等である。このように検索条件に、ファイルのプロパティ情報を加えることで、例えば作成者等を特定したい場合の類似文書の絞り込みが適切にできる。

#### 【0019】

検索要求元の検索条件指定部26はクライアント12のWWWブラウザ16で提供され、WWWブラウザ16の検索要求画面で指定したファイル内容をネットワークを介してWWWサーバ18に送信して検索マシン20に引き渡す。

#### 【0020】

本発明は、またサーバ等の検索側の文書情報検索装置となる検索マシン20を提供する。この検索マシン20としての文書情報検索装置は、検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存している検索データベース22、文書ファイルを検索条件に指定したネットワークからの検索要求によって受信したファイル内容からテキスト文を抽出するテキスト抽出処理部36、テキスト文の形態素解析により名詞を抽出する形態素回析部28、名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部40、及びキーワードによる検索データベースの検索で類似する文書を検索して要求元に検索結果を通知する検索実行部42を備える。

#### 【0021】

本発明は、クライアント等の検索要求元からのネットワークを経由した検索要求に基づいてサーバ等の検索マシン側で文書情報を検索して応答する文書情報検索方法を提供する。この文書情報検索方法は、

検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎にサーバの検索データベースに保存し；

文書ファイルを検索条件に指定した場合に、指定したファイル内容を検索要求と共にネットワークを経由して検索側に送信し；

検索側で、検索要求に伴って受信したファイル内容からテキスト文を抽出すると共にテキスト文の形態素解析により名詞を抽出し、次に名詞の中から重要語を抽出して論理和でつなげたキーワードを生成し、該キーワードによる検索データベースの検索で類似する文書を検索してクライアントに検索結果を通知すること

を特徴とする。この文書情報検索方法の詳細は装置構成と基本的に同じになる。

#### 【0022】

更に、本発明は、文書情報検索プログラムを格納したコンピュータ可読の記録媒体を提供するもので、この文書情報検索プログラムは、文書ファイルを検索条件に指定した検索要求を受信するステップと、検索要求に伴って受信したファイル内容からテキスト文を抽出するステップと、テキスト文の形態素解析により名詞を抽出するステップと、名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するステップと、キーワードによるデータベースの検索で類似する文書を検索して要求元に検索結果を通知するステップとを備える。

#### 【0023】

##### 【発明の実施の形態】

図2は、本発明による文書情報検索装置のシステム構成であり、インターネットやイーサネットを利用したサーバクライアント型の検索システムとして構築した場合を例にとっている。

#### 【0024】

図2において、サーバ10に対しては、ユーザ側のクライアント12がインターネット／インターネット14を介して接続される。クライアント12には検索用のWWWブラウザ16が設けられており、このWWWブラウザ16を利用してサーバ10に対し文書情報の検索要求を行い、サーバ10側の検索結果を表示する。

#### 【0025】

サーバ10には、WWWサーバ18、検索マシン20、文書データベース24が設けられている。検索マシン20には検索データベース22が格納されている。また文書データベース24には検索対象文書25が格納されている。更にWWWサーバ18に対して外部の文書管理サーバ44、48が接続され、この文書管理サーバ44、48にも文書データベース46、50が設けられており、それぞれ検索対象文書25を格納している。

#### 【0026】

サーバ10に設けているWWWサーバ18は、ブラウザ16からの検索要求を受信して検索マシン20に対し検索を依頼する。また検索マシン20から返ってきた検索結果をブラウザ16に返して表示させる。

#### 【0027】

検索データベース22は、全文検索を高速に処理するために、検索対象となる文書に記述されている重要な単語の集合で作られたインデックスを管理する保管庫として機能する。このインデックスには文書の文書名やその保管場所が記録されており、ブラウザ16から検索要求を受けた際には、検索データベース22のインデックスを対象に検索マシン20が検索処理を実行する。

#### 【0028】

文書データベース24には、文書管理サーバ44, 48から収集した検索対象文書25が格納されており、この文書データベース検索対象文書25を対象に検索データベース22のインデックスが作成されている。

#### 【0029】

このようなサーバクライアント型の検索システムにあっては、クライアント12のブラウザ16を使用して、ユーザが指定した検索条件をインターネット／イントラネット14を経由してサーバ10側のWWWサーバ18に送る。WWWサーバ18で受信された検索要求に含まれる指定された検索条件が、WWWサーバ18から検索マシン20に送られる。

#### 【0030】

検索マシン20は検索データベース22から検索条件にあった文書を検索し、検索結果をWWWサーバ18に通知する。WWWサーバ18は検索マシン20からの検索結果をクライアント12のブラウザ16に送って表示させる。

#### 【0031】

ユーザはブラウザ16で処理された検索結果を見て、検索結果に記述されたリンクを選択することで、選択された文書の中からユーザが希望する検索対象文書25をWWWサーバ18経由でアップロードして内容を見ることができる。

#### 【0032】

図3は図2の検索システムにおける機能構成のブロック図である。まずユーザ

側となるWWWブラウザ16には検索条件指定部26が設けられている。本発明の検索条件指定部26は、検索条件としてユーザがインターネットや電子メールなどで入手した文書ファイルを直接、検索条件として指定し、指定したファイル内容をインターネット／インターネット14経由でWWWサーバ18を経由して検索マシン20のファイル指定検索部30に送信する。

#### 【0033】

また検索条件指定部26は、本発明で新たに提供されるファイル指定の検索条件とする以外に、

- (1) キーワード検索、
- (2) 文書のタイトル、作成者、本文ごとにキーワードを指定して検索する詳細検索、
- (3) 日常的な言葉や文章を入力することにより本文内容を関連する文書を検索する文章検索、更に、
- (4) 検索データベース22に登録済みの既存文書を検索条件に使用した類似文書検索、

などの検索条件の指定も可能である。

#### 【0034】

WWWサーバ18側に設けられた検索マシン20には、検索データベース作成部28、文書検索部30及び文書参照部32が設けられている。検索データベース作成部28は検索データベース22にインデックスを作成して登録する。

#### 【0035】

即ち検索データベース作成部28は、文書データベース24に収集されて保存されている検索対象文書25の1つ1つについて、検索対象文書25に記述されている重要語を抽出し、抽出された単語の集合で構成されたインデックスを作成して保存する。もちろん、このインデックスには検索対象文書の文書名や保管場所などが併せて記録されている。

#### 【0036】

文書検索部30は、WWWブラウザ16の検索条件指定部26から送信された検索条件としてファイルを指定した際のファイル内容からキーワードを生成し、

検索データベース22のインデックスに含まれている重要単語の集合との検索照合を行い、WWWブラウザ16で検索条件として指定したファイルの文書に類似する文書を検索し、検索結果をWWWサーバ18からWWWブラウザ16に返して表示させる。

#### 【0037】

文書参照部32は、WWWブラウザ16で送出された検索結果としての文書一覧から参照したい文書を選択すると、WWWサーバ18を介して文書参照部32に通知されると、文書データベース24の中から要求された参照文書を取り出してWWWブラウザ16に返す。

#### 【0038】

図4は、図3の検索マシン20に設けた本発明の文書検索部30の機能構成の詳細である。

#### 【0039】

図4において、文書検索部30には、検索指定ファイル格納部34、テキスト抽出処理部36、形態素解析部38、キーワード作成部40及び検索実行部42が設けられている。また検索データベース22内には、図3の検索データベース作成部28で作成された文書データベース24内の検索対象文書25のそれぞれの重要な単語の集合、文書名、保管場所などで構成されたインデックス52が格納されている。

#### 【0040】

文書検索部30の検索指定ファイル格納部34には、図3のWWWブラウザ16における検索条件指定部26のファイル指定により送信されたファイル内容が格納される。

#### 【0041】

ここでWWWブラウザ16側からのファイル内容の転送は、検索条件として指定した文書ファイルの先頭ファイル部分、例えば先頭の1KBを切り出してWWWサーバ18側に検索要求と共に送信する。

#### 【0042】

このように検索条件として送信するファイル容量を例えば1KBというように

固定容量とすることで、検索条件として指定している文書ファイルのサイズの大小に関わらず、検索マシン20側に対する文書内容の転送負荷を一定にし、また検索マシン20におけるファイル指定部検索部30による検索処理の安定化と迅速化を図る。

#### 【0043】

テキスト抽出処理部36は、検索指定ファイル格納部34に格納された検索条件として指定されたファイル内容からテキスト文書を抽出する。WWWブラウザ16における検索条件として指定される文書ファイルの形式としては、電子メールのテキストファイル、インターネットにおけるHTMLファイル、更には集計リストのエクセルファイルなどの様々なファイル形式があることから、これらのファイル形式の相違に対して検索機能を提供可能とするため、各種の形式の文書ファイルの中からテキスト抽出処理部36によりテキスト文書のみを抽出して検索条件に使用するようにしている。

#### 【0044】

続いて設けた形態素解析部38は、抽出されたテキスト文書の中に含まれる名詞を形態素解析を用いて抽出する。形態素解析部38で抽出された文書内容の中の名詞はキーワード作成部40に送られ、キーワード作成部40においては重要な名詞をキーワード作成のために抽出する。

#### 【0045】

キーワード作成部40における重要語の抽出は、まず各名詞が検索データベース22のインデックス52の中に登録している文書数Nの内の何文書で出現するかの出現数Hのカウントを行う。

#### 【0046】

そして、インデックス52中における文書出現数Hが求められたならば、出現数Hが予め定めた範囲内、例えば

$$(2N/3) \geq H \geq 1$$

となる出現数の単語を選択する。このように選択された単語の内の出現数Hが大きい上位10個の単語をキーワード作成のために選択する。そして選択した重要単語10個を論理和で繋げたクエリ式を作成して検索実行部42に提供する。

## 【0047】

検索実行部42はキーワード作成部40から与えられたクエリ式に基づいて検索データベース22のインデックス52との検索照合を行い、所定の類似度を満たすインデックスを検索結果として抽出し、検索結果をWWWサーバ18によりWWWブラウザ16側に送信し、検索結果の文書一覧の形でユーザに参照できるようとする。

## 【0048】

更に文書検索部30にあっては、検索指定ファイル格納部34に格納された検索条件として指定されたファイルのプロパティ情報を利用した文書検索もできる。このためWWWブラウザ16の検索条件指定部26は、検索条件として文書ファイルを指定した際に、指定した文書ファイルのプロパティ情報を抽出し、検索条件として指定した文書の先頭ファイル部分、例えば先頭ファイル部分1KBと共にプロパティ情報を検索マシン20側に送信する。

## 【0049】

図14の文書検索部30にあっては、ファイル内容からのテキスト文の抽出、形態素解析による名詞抽出、名詞について重要語の選択によるキーワード作成に加え、検索指定ファイル格納部34に格納されているファイル内容に付加されたプロパティ情報から例えば作成日や作成者、題名などを抽出し、キーワード作成部40でプロパティ情報をキーワードに含め、検索実行部42で検索データベース22のインデックス52の検索を行う。

## 【0050】

図5は、図3の検索マシン20に設けている検索データベース作成部28によるインデックス作成処理の説明図である。この検索データベース作成部28にあっては、ロボット54が外部の文書データベース46、50から文書66を収集してテンポラリファイル62に格納し、同時に収集文書リストファイル64に収集した文書66のリストを加える。

## 【0051】

続いてロボット54はテキスト抽出部56に処理を渡し、テキスト抽出部56は収集文書リストファイル64から収集文書66を取り出し、抽出テキストファ

イル68に格納する。

#### 【0052】

次に重要語抽出部58に処理を渡し、重要語抽出部58は抽出テキストファイル68の該当テキスト文書の中から形態素解析により名詞を抽出し、名詞についてそれぞれ出現頻度をカウントし、例えば出現頻度の高い単語の上位10個を重要語として抽出して重要語ファイル70に格納する。

#### 【0053】

次にインデックス作成部60に処理を渡し、インデックス作成部60は重要語ファイル70から、その文書について例えば上位10個の重要語の集合を取り出し、更に文書名と保管場所を加えたインデックスを作成し、検索データベース22にインデックス情報として保存する。

#### 【0054】

図6は、図3のWWWブラウザ16による検索条件の指定と検索結果の表示を行うブラウザ処理のフローチャートである。ユーザがWWWブラウザ16の検索機能を開くと、ステップS1で検索画面が表示され、この検索画面を表示して、ステップS2で文書ファイルを指定した検索条件の指定操作を行う。

#### 【0055】

続いてステップS3で検索起動の有無をチェックしており、検索起動を判別すると、ステップS4でファイル指定検索か否かチェックする。ファイル指定検索であればステップS5に進み、ユーザが指定したファイルを読み出し、ステップS6で指定ファイルの先頭1KBを検索要求メッセージと共にサーバに送信する。

#### 【0056】

ファイル指定検索でなければ、ステップS7で、それ以外の検索例えばキーワード検索に対応した検索要求メッセージをサーバに送信する。ステップS6で指定ファイルの先頭部分をサーバに送信すると、ステップS8で検索結果の受信待ちとなる。

#### 【0057】

ステップS8でサーバから検索結果が受信されると、ステップS9に進み、検

索結果の表示操作処理を行ってユーザは検索内容を見る。このようなステップS1～S9の処理を、ステップS10で検索画面を閉じる検索終了指示があるまで繰り返す。

#### 【0058】

図7は、図6のブラウザ処理において検索条件として文書ファイルを指定した場合の具体的な手順と画面の様子を表わしている。

#### 【0059】

図7において、まずユーザは検索条件に指定しようとする文書ファイル72を例えばインターネットから取得している。そしてユーザは文書ファイル72の内容を見て、この文書ファイル72に類似する文書検索を行うため、文書ファイル72の内容を予め指定したファイル、例えばファイル「news.txt」に保存する。

#### 【0060】

続いてユーザはキーワード入力画面74を開く。キーワード入力画面74にはキーワード入力部76、ファイル指定部78、参照ボタン80及び検索実行ボタン82が設けられている。そこで、ユーザがキーワード入力画面74の参照ボタン80を押すことでファイル選択ダイアログ84を表示する。

#### 【0061】

このファイル選択ダイアログ84の中には、検索条件として指定したい文書ファイル72が保存されていることから、ファイル名「news.txt」をマウスクリックして選択すると、キーワード入力画面74のファイル指定部78に選択したファイル名「news.txt」が設定される。

#### 【0062】

このようにしてファイル指定部78によるファイル指定が済んだならば、検索実行ボタン82を押すことで、検索条件として指定された文書ファイル「news.txt」の文書内容の先頭1KBが検索要求と共にサーバに対し送信される。

#### 【0063】

図8は、図4の文書検索部30によって実現されるサーバ検索処理のフロー

ヤートである。このサーバ検索処理は、ステップS1で検索条件として指定された文書ファイルを読み込み、ステップS2で文書ファイルからテキスト文書の抽出処理を行う。次にステップS3で、抽出したテキスト文書の内容について形態素解析を用いて名詞を抽出する。次にステップS4で、名詞として抽出した各単語が検索データベース22に設けているインデックス52の中の文書数N内の何文書に出現するかの出現数Hのカウント処理を行う。

#### 【0064】

各単語のインデックス中の出現数Hがカウントできたならば、ステップS5で出現数Hが（ $2N/3$ ）以下で1以上となる範囲の単語をまず選択し、この選択した単語のうち出現数Hが大きい上位10個の単語をキーワードに使用する重要語として選択する。続いてステップS6で、重要語として選択した10個の単語を論理和で繋げたクエリ式を生成する。

#### 【0065】

そしてステップS7で、検索キーワードとして生成されたクエリ式による検索データベースのインデックスの検索を行い、生成したキーワードに対し所定の類似度を持つインデックスの内容を検索文書として一覧表にまとめ、ステップS8で検索結果をブラウザに送信する。

#### 【0066】

図9は、図8のステップS2のテキスト抽出処理の詳細である。このテキスト抽出処理にあっては、ステップS1で文書ファイルの拡張子を解読する。ファイル拡張子からステップS2でHTML文書であることが認識されると、ステップS3に進み、HTML文書におけるボディタグ内のデータをテキストデータ本文として抽出し、タグデータは取り除く。

#### 【0067】

例えば図10(A)のようなHTMLファイルを例にとると、<>で挟まれたボディ単語の中のデータをテキストデータ本文として取り出して、このタグデータは取り除くことで、図10(B)のような抽出テキスト文書が得られる。

#### 【0068】

次にステップS4で、OSで管理しているファイルのプロパティ情報を獲得す

る。このプロパティ情報は、例えばファイル所有者や文書タイプなどを含んでいる。

#### 【0069】

図11は、インターネットから入手した文書ファイルのプロパティ情報の例であり、このプロパティ情報にあっては文書タイトル「文書管理システムについて」や作成日、変更日などが存在し、これらのプロパティデータをキーワード生成のために獲得する。

#### 【0070】

一方、ステップS2でHTML文書ではなく例えばエクセル文書などであった場合には、ステップS5で文書ライブラリにファイルを渡し、テキストデータを獲得する。続いてステップS6で、プロパティ情報獲得関数により文書ごとに設定されているファイルプロパティ情報例えば作成者や文書タイトルなどを獲得する。

#### 【0071】

図12は本発明で検索条件として指定するHTMLファイル以外のファイルとしてエクセルファイルを示している。この図12のエクセルファイルについて、文書ライブラリに渡してテキストデータを獲得すると、図13の抽出テキスト文書に示すようなエクセル文書中に書き込まれているテキスト文書を抽出した結果が得られる。

#### 【0072】

このようなテキスト抽出処理で得られたHTML文書やエクセル文書からのテキスト文書、更にはプロパティ情報から得られたテキスト文書をひとまとめにして、図8のステップS3で形態素解析を用いて名詞を抽出し、ステップS4、S5で、データベースのインデックスの参照で重要語の上位10個をキーワードに選択してクエリ式を作り、データベースのインデックス検索を行って検索結果を得ることができる。

#### 【0073】

尚、図9のテキスト抽出処理におけるステップS4、S6のプロパティ情報の獲得は、WWWブラウザ16におけるユーザ側の指定によってプロパティ情報を

使用するか否かの選択が可能であり、プロパティ情報を使うか否かは検索結果をどの程度絞り込むかのユーザ判断に依存する。

## 【0074】

本発明はまた、図4の検索マシン20に文書検索部30の処理機能を実行する文書情報検索プログラムを記録したコンピュータ読取り可能な記録媒体を提供する。この記録媒体の実施形態としては、CD-ROMやフロッピディスクなどのリムーバブルな可搬型記録媒体、回線によりプログラムを提供するプログラム提供者の記憶装置、更にプログラムをインストールした処理装置のRAMやハードディスクなどのメモリ装置を含む。

## 【0075】

また記録媒体によって提供された図4の文書検索部30の機能を実現する文書情報検索プログラム、具体的には図8及び図9のフローチャートの処理を実行するステップを備えた文書情報検索プログラムは、サーバなどの処理装置にローディングされ、その主メモリ上で実行される。

## 【0076】

またサーバ側にローディングされた本発明の文書情報検索プログラムは、クライアント側からサービス要求を受けると、クライアント12側にファイル指定による検索条件の指定を行うWWWブラウザ機能をアップロードし、ユーザによる検索システムの利用を可能とする。

## 【0077】

尚、上記の実施形態はサーバクライアント型の検索システムを例にとるものであったが、本発明はこれに限定されず、ホスト端末型や適宜のシステム形態をとることができる。また本発明は上記の実施形態に限定されず、その目的と利点を損なわない適宜の変形を含む。更にまた本発明は上記の実施形態に示した数値による限定は受けない。

(付記)

(付記1)

ネットワークを経由した検索要求に基づいて文書情報を検索して応答する文書情

報検索装置に於いて、

検索要求元に、検索条件としてファイルを指定し、指定したファイル内容をネットワークを経由して送信する検索条件指定部を設け、

検索側に、前記検索条件指定部から送信されたファイル内容からキーワードを生成してデータベースから類似文書を検索する文書検索部を設けたことを特徴とする文書情報検索装置。 (1)

(付記2)

付記1記載の文書情報検索装置に於いて、前記検索条件指定部は、指定されたファイル内容の先頭ファイル部分を送信することを特徴する文書情報検索装置。

【0078】

(付記3)

付記1記載の文書情報検索装置に於いて、前記検索条件指定部は検索条件として指定するファイルにHTMLファイル及びエクセルファイルを含むことを特徴とする文書情報検索装置。

【0079】

(付記4)

付記1記載の文書情報検索装置に於いて、

前記データベースは、検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存し、

サーバの文書検索部は、

検索要求に伴って受信したファイル内容からテキスト文を抽出するテキスト抽出処理部と、

前記テキスト文の形態素解析により名詞を抽出する形態素回析部と、

前記名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキーワード生成部と、

前記キーワードによる検索データベースの検索で類似する文書を検索してクライアントに検索結果を通知する検索実行部と、

を備えたことを特徴とする文書情報検索装置。 (2)

(付記5)

付記4記載の文書情報検索装置に於いて、前記キーワード生成部は、各名詞が前記文書データベースに格納した検索文書毎のインデックス中の何文書に出現するかの出現数をカウントし、所定範囲の出現数をもつ上位の所定数の単語を選択してキーワードを生成することを特徴とする文書情報検索装置。（3）

（付記6）

付記5記載の文書情報検索装置に於いて、前記キーワード生成部は、インデックス中の文書数Nとした場合、出現数Hが $2N/3 \geq H \geq 1$ の範囲の出現数をもつ上位の10個の単語を選択してキーワードを生成することを特徴とする文書情報検索装置。（4）

（付記7）

付記3記載の文書情報検索装置に於いて、前記キーワード生成部は検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索させることを特徴とする文書情報検索装置。（5）

（付記8）

付記7記載の文書情報検索装置に於いて、前記プロパティ情報は、検索要求に伴って受信したファイルの作成者、文書タイトル等であることを特徴とする文書情報検索装置。

【0080】

（付記9）

付記1記載の文書情報検索装置に於いて、前記検索要求元の検索条件指定部はクライアントのWWWブラウザで提供され、前記WWWブラウザの検索要求画面で指定したファイル内容をネットワークを介してWWWサーバの検索マシンに送信して前記ファイル指定検索部に引き渡すことを特徴とする文書情報検索装置。

【0081】

（付記10）

検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎に保存しているデータベースと、

文書ファイルを検索条件に指定したネットワークからの検索要求によって受信したファイル内容からテキスト文を抽出するテキスト抽出処理部と、

前記テキスト文の形態素解析により名詞を抽出する形態素回析部と、  
前記名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するキー  
ワード生成部と、  
前記キーワードによるデータベースの検索で類似する文書を検索して要求元に検  
索結果を通知する検索実行部と、  
を備えたことを特徴とする文書情報検索装置。 (6)

(付記11)

付記10記載の文書情報検索装置に於いて、前記キーワード生成部は、各名詞が  
前記文書データベースに格納した検索文書毎のインデックス中の何文書に出現す  
るかの出現数をカウントし、所定範囲の出現数をもつ上位の所定数の単語を選択  
してキーワードを生成することを特徴とする文書情報検索装置。

【0082】

(付記12)

付記10記載の文書情報検索装置に於いて、前記データベースにインデックス情  
報と共に検索対象文書から抽出したプロパティ情報を保存し、前記キーワード生  
成部は検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キ  
ーワードに含めて検索することを特徴とする文書情報検索装置。 (7)

(付記13)

ネットワークを経由した検索要求に基づいて文書情報を検索して応答する文書情  
報検索方法に於いて、

検索対象文書から抽出した重要語を列挙したインデックス情報を文書毎にデータ  
ベースに保存し、

前記検索要求元で検索条件にファイルを指定した場合に、指定したファイル内容  
を検索要求と共にネットワークを経由してサーバに送信し、

検索側で、検索要求に伴って受信したファイル内容からテキスト文を抽出すると  
共にテキスト文の形態素解析により名詞を抽出し、次に名詞の中から重要語を抽  
出して論理和でつなげたキーワードを生成し、該キーワードによるデータベース  
の検索で類似する文書を検索して検索結果を応答することを特徴とする文書情報  
検索方法。 (8)

(付記14)

付記13記載の文書情報検索方法に於いて、前記キーワードの生成として、各名詞が前記データベースに格納した文書毎のインデックス中の何文書に出現するかの出現数をカウントし、所定範囲の出現数をもつ上位の所定数の単語を選択してキーワードを生成することを特徴とする文書情報検索方法。

【0083】

(付記15)

付記14記載の文書情報検索方法に於いて、検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索することを特徴とする文書情報検索方法。(9)

(付記16)

文書ファイルを検索条件に指定した検索要求を受信するステップと、検索要求に伴って受信したファイル内容からテキスト文を抽出するステップと、テキスト文の形態素解析により名詞を抽出するステップと、名詞の中から重要語を抽出して論理和でつなげたキーワードを生成するステップと、前記キーワードによるデータベースの検索で類似する文書を検索して要求元に検索結果を通知するステップと、を備えた検索プログラムを格納したコンピュータ可読の記録媒体。(10)

(付記17)

付記16記載の記録媒体に於いて、前記文書情報検索プログラムのキーワードを生成するステップは、各名詞が前記データベースに格納した文書毎のインデックス中の何文書に出現するかの出現数をカウントし、所定範囲の出現数をもつ上位の所定数の単語を選択してキーワードを生成することを特徴とする記録媒体。

【0084】

(付記18)

付記14記載の記録媒体に於いて、前記文書情報検索プログラムは、更に検索要求に伴って受信したファイルから抽出したプロパティ情報を前記キーワードに含めて検索するステップを備えたことを特徴とする記録媒体。

## 【0085】

## 【発明の効果】

以上説明してきたように本発明によれば、ユーザが電子メールやインターネットなどで興味のある内容を含む文書を入手した際に、この文書に類似した内容の文書検索を文書ファイルを直接検索条件として指定することで、内容が類似する文書を簡単且つ素早く検索することができ、手間の掛かる文書内容に基づいたキーワードの入力を不要とし、ユーザによる類似文書の探し出しが極めて効率的に実現できる。

## 【0086】

またファイル指定による文書検索に必要なキーワードの生成において、文書内容から重要な単語を抽出する以外に、文書ファイルの持っているプロパティ情報からも重要な単語を抽出してキーワードに含めることで、データベースに登録している既存文書の類似検索の絞り込みが、より適切に行われ、検索の精度を高めることができる。

## 【図面の簡単な説明】

## 【図1】

本発明の原理説明図

## 【図2】

本発明のシステム構成の説明図

## 【図3】

本発明の機能構成のブロック図

## 【図4】

本発明による文書検索部のブロック図

## 【図5】

図3の検索データベース作成部の処理説明図

## 【図6】

図3のブラウザ処理のフローチャート

【図7】

本発明の検索条件に文書ファイルを指定する検索要求操作の説明図

【図8】

本発明のサーバ検索処理のフローチャート

【図9】

図8のテキスト抽出処理のフローチャート

【図10】

図8の処理によりHTMLファイルからのテキスト文書を抽出する説明図

【図11】

本発明の検索に使用するHTMLファイルに設けたプロパティ情報の説明図

【図12】

図8の処理によりテキスト抽出対象とするExcel文書の説明図

【図13】

図12のExcel文書から抽出したテキスト文書の説明図

【符号の説明】

10 : サーバ

12 : クライアント

14 : インターネット／インターネット

16 : WWWブラウザ

18 : WWWサーバ

20 : 検索マシン

22 : 検索データベース

24, 46, 50 : 文書データベース

25 : 検索対象文書

26 : 検索条件指定部

28 : 検索データベース作成部

30 : 文書検索部

32 : 文書参照部

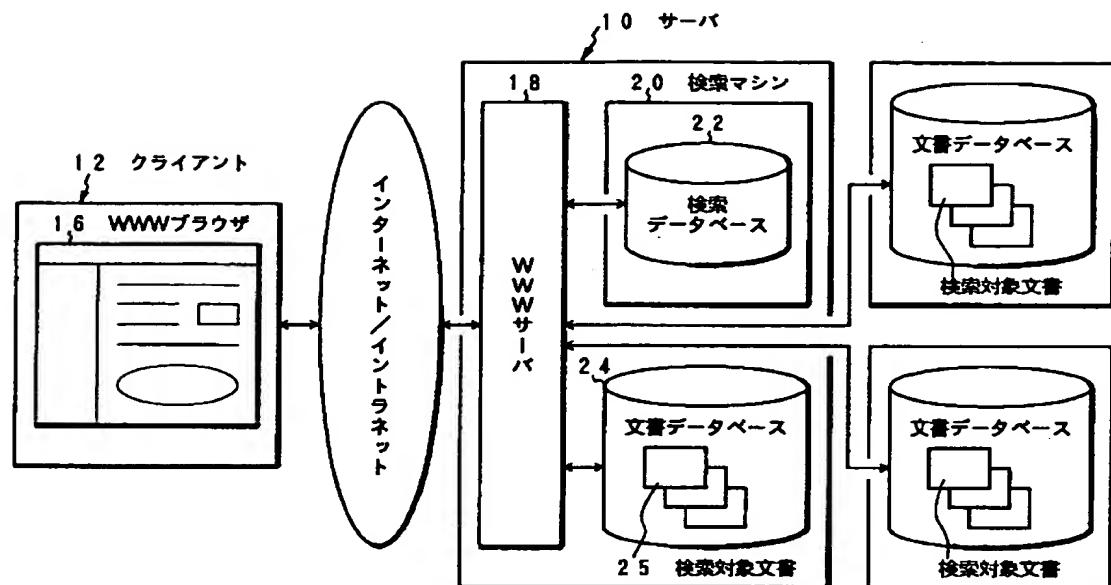
34 : 検索指定ファイル格納部  
36 : テキスト抽出処理部  
38 : 形態素解析部  
40 : キーワード作成部  
42 : 検索実行部  
44, 48 : 文書管理サーバ  
54 : ロボット  
56 : テキスト抽出部  
58 : 重要語抽出部  
60 : インデックス作成部  
62 : テンポラリファイル  
64 : 収集文書リストファイル  
66 : 文書  
68 : 抽出テキストファイル  
70 : 重要語ファイル

【書類名】 図面

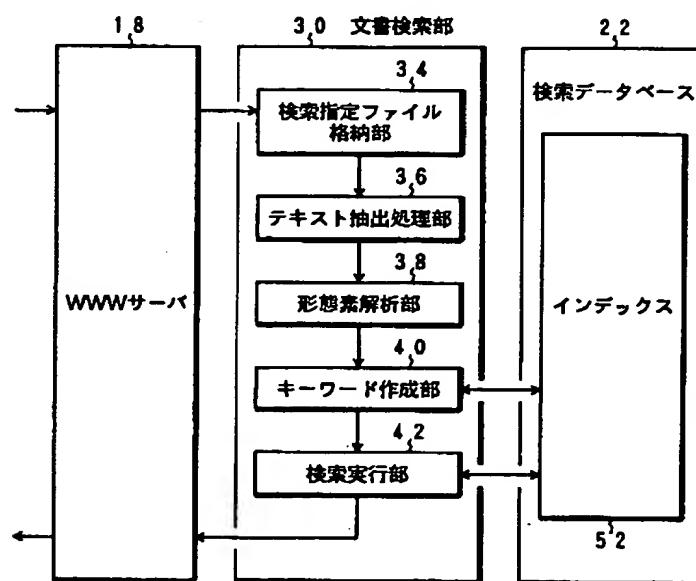
【図1】

本発明の原理説明図

(A)

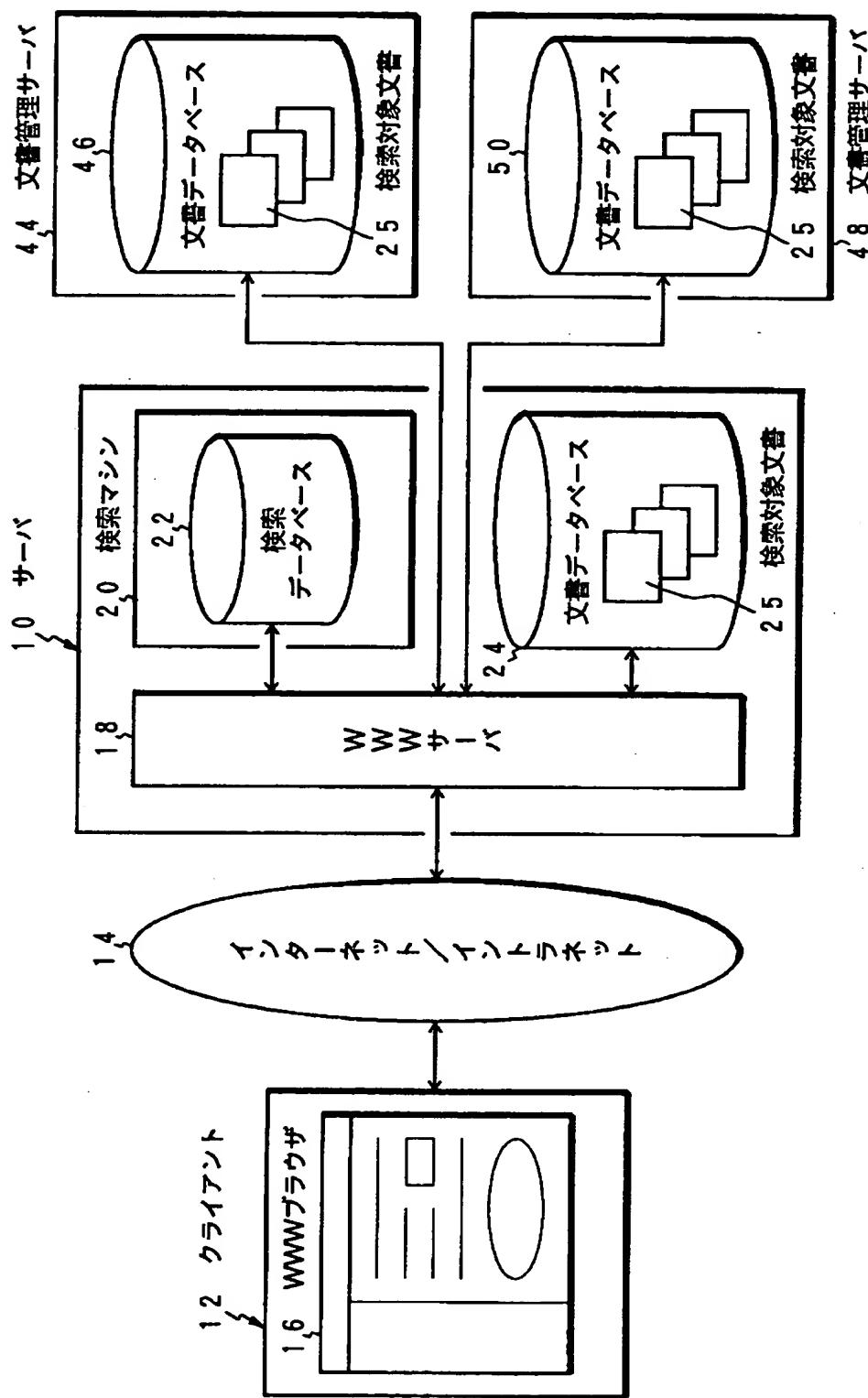


(B)



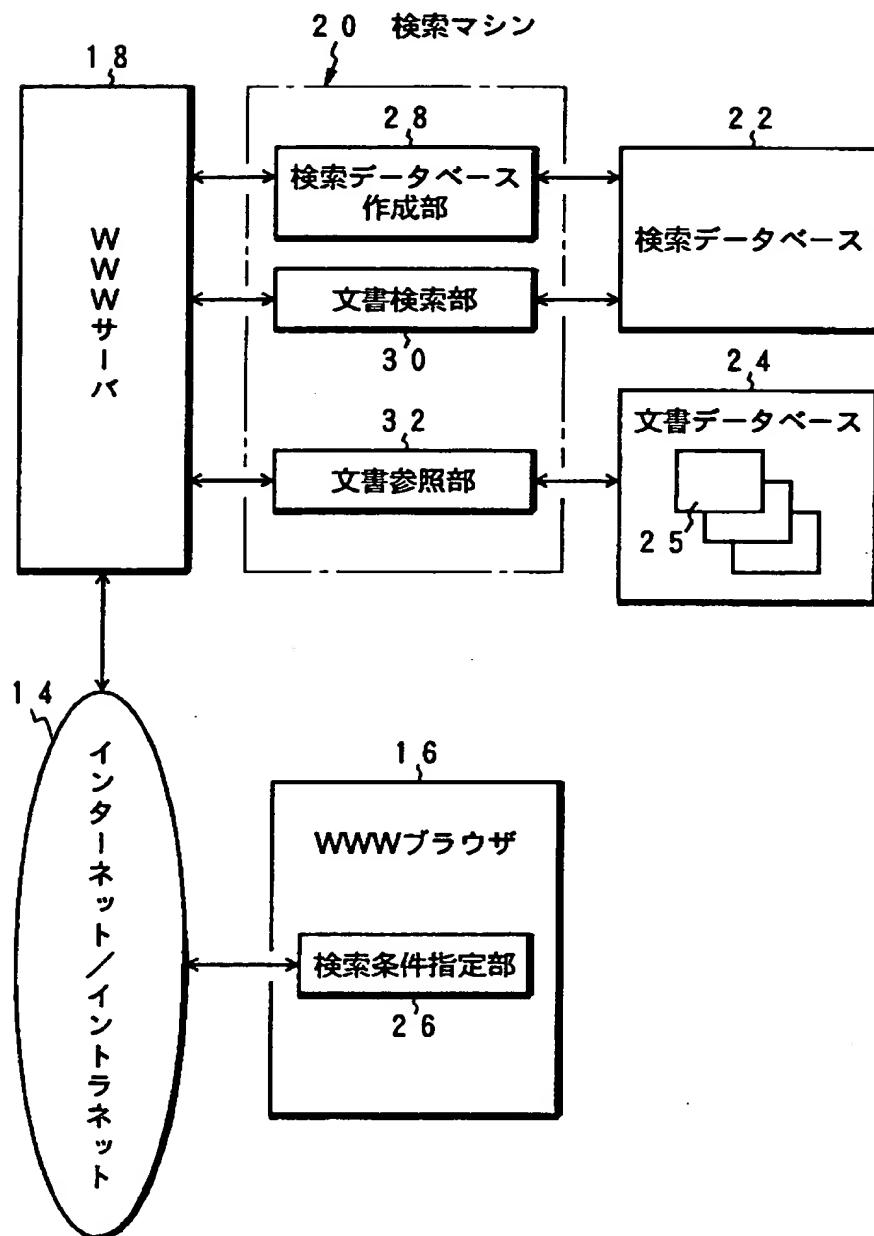
【図2】

## 本発明のシステム構成の説明図



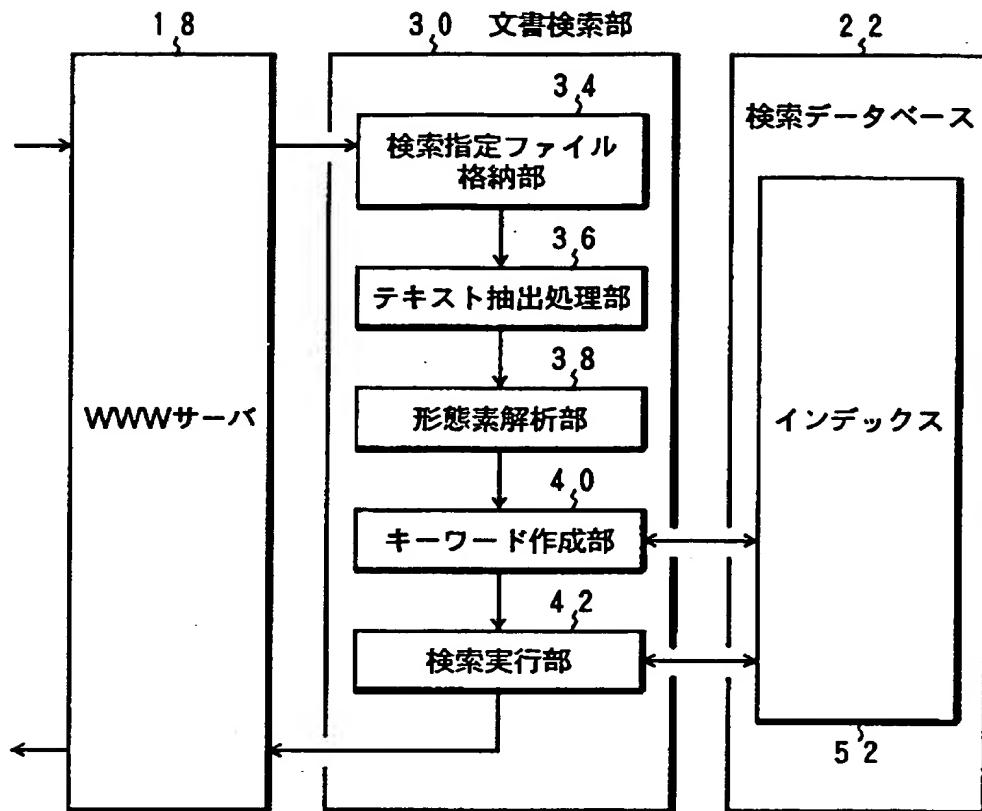
【図3】

## 本発明の機能構成のブロック図

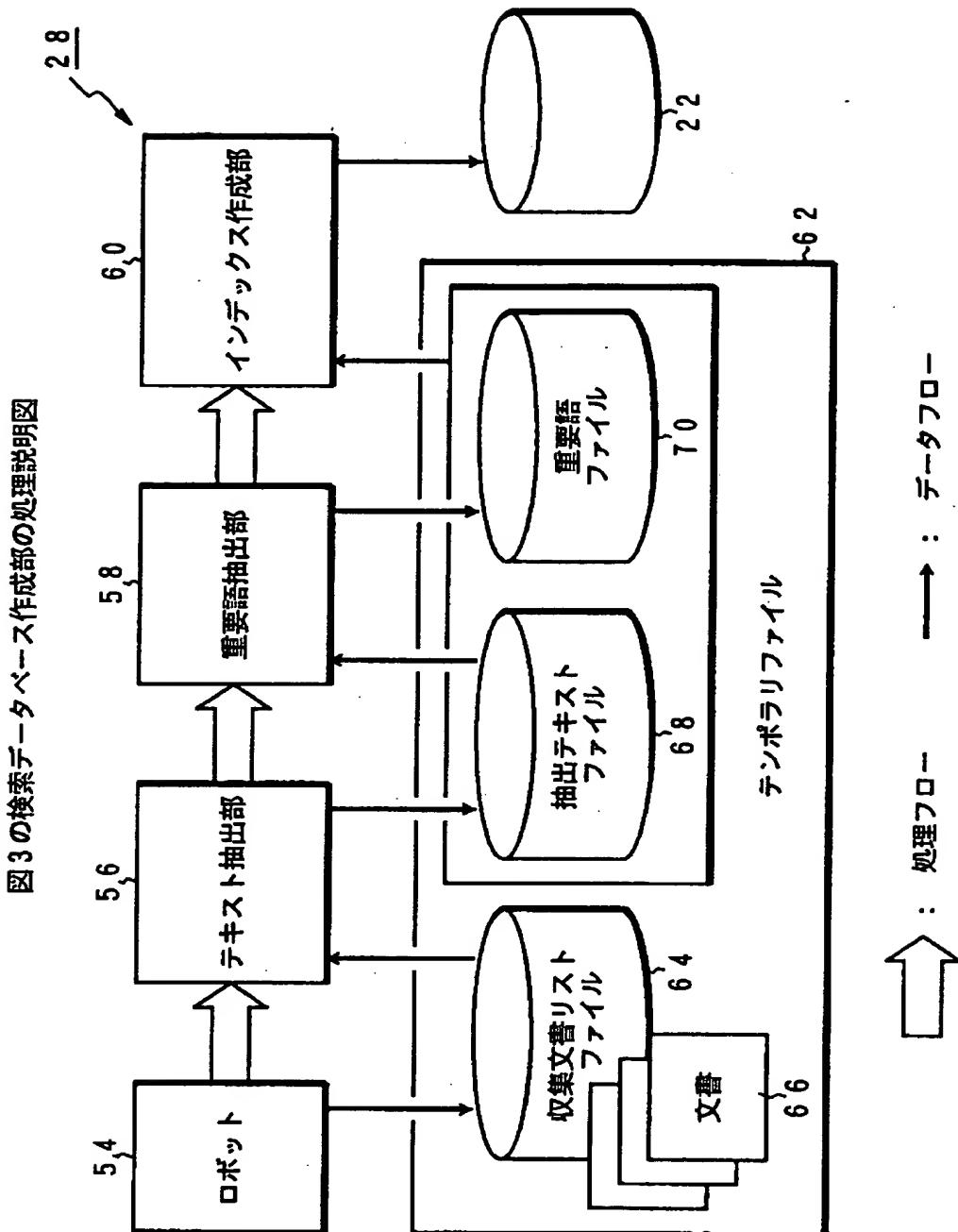


【図4】

## 本発明による文書検索部のブロック図

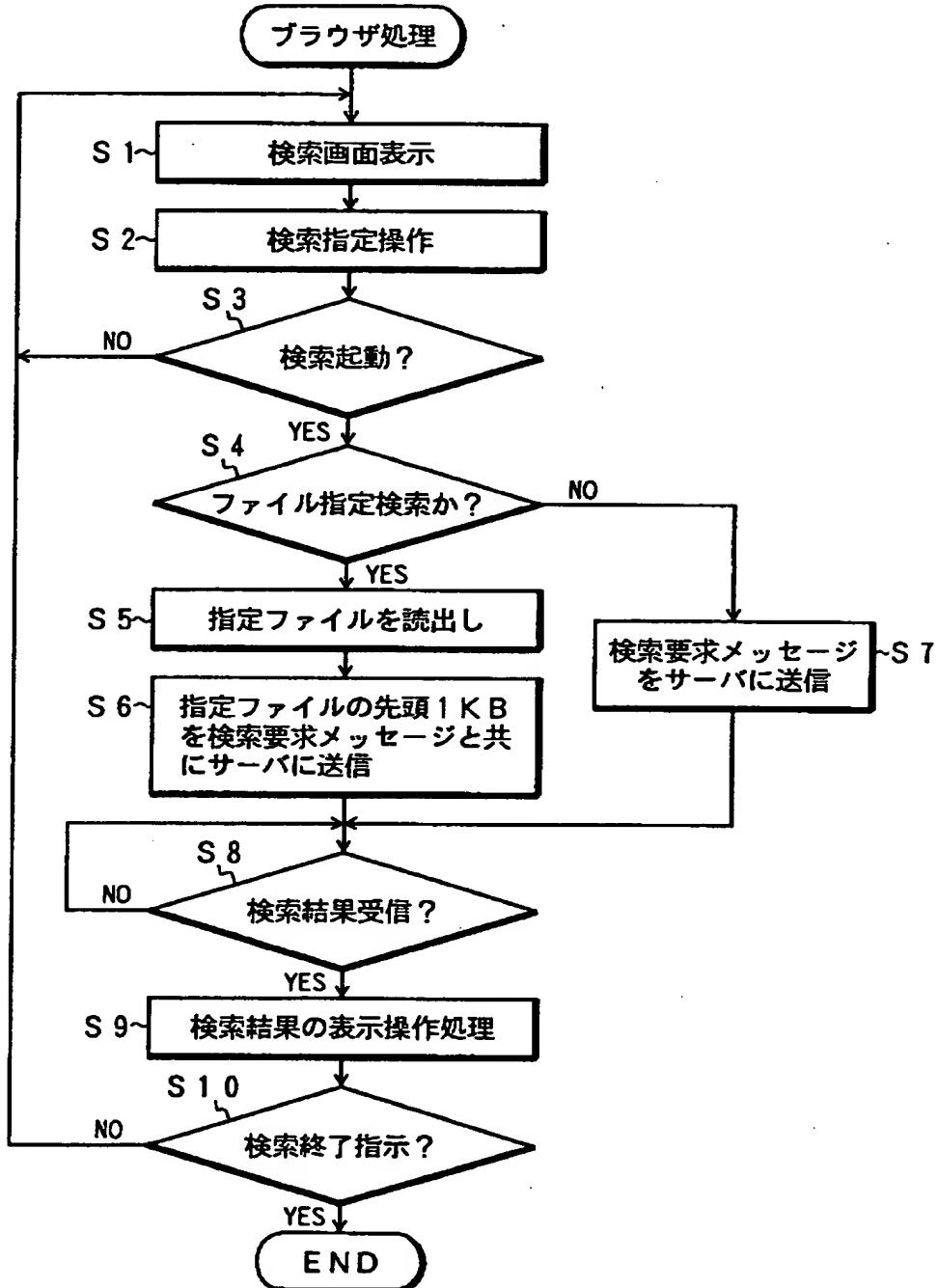


【図5】

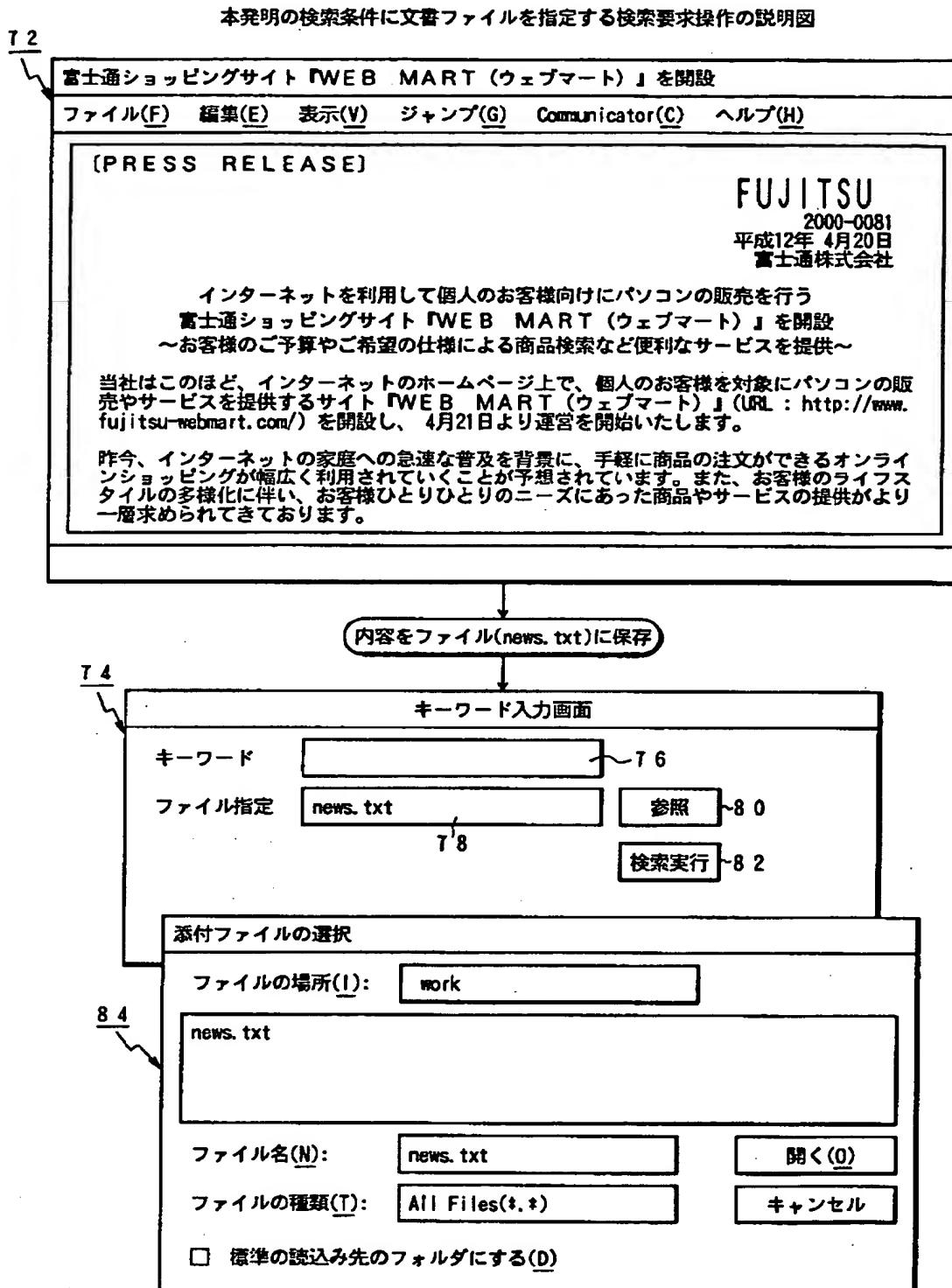


【図6】

図3のブラウザ処理のフローチャート

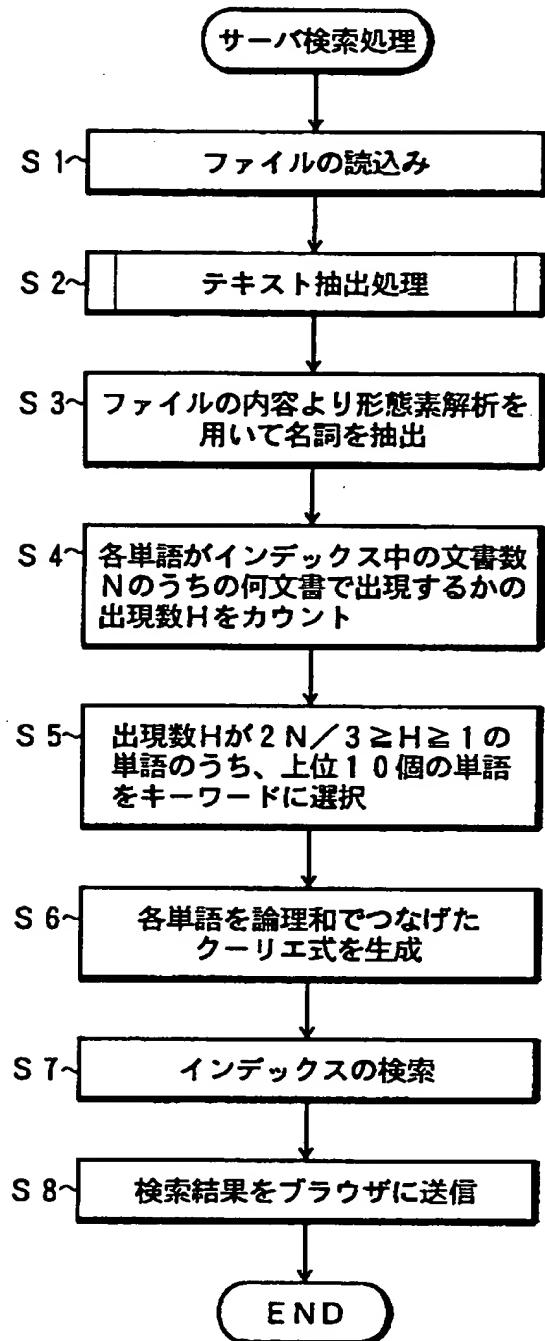


【図7】



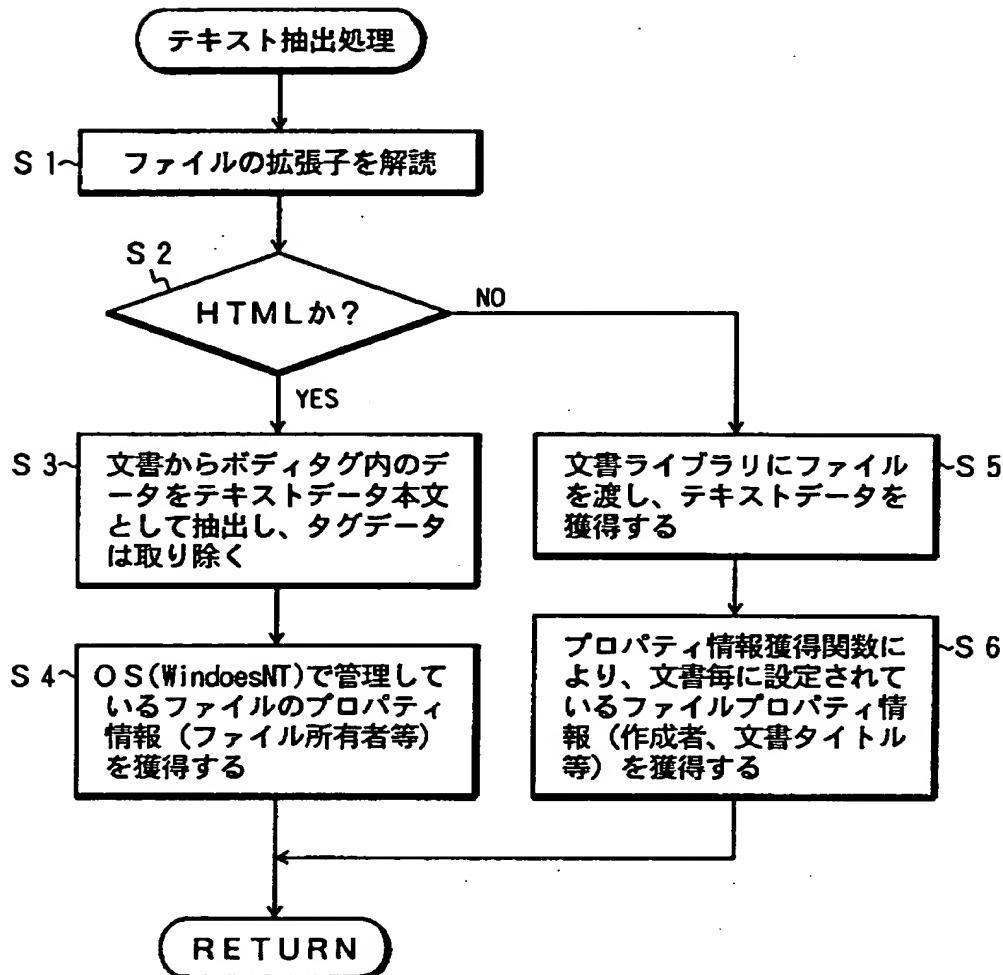
【図8】

## 本発明のサーバ検索処理のフローチャート



【図9】

図8のテキスト抽出処理のフローチャート



## 【図10】

図8の処理によりHTMLファイルからのテキスト文書を抽出する説明図

(A)

## HTMLファイル

```
<HTML>
<HEAD>
<TITLE>検索データベース</TITLE>
</HEAD>
<BODYBGCOLOR="#FFFFFF">
<H4>検索データベース</H4>
<HR>

<P>検索に使用するデータベースのことです。<BR>
検索対象の文書のタイトルや保管場所が記録されています。検索データベースは定期的に
更新され、最新の情報が記録されます。IntelligentSearch は検索データベースにより、
高速検索を実現しています。<BR>

</BODY>
</HTML>
```

(B)

## 抽出テキスト文書

検索データベース検索データベース検索に使用するデータベースのことです。検索対象の  
文書のタイトルや保管場所が記録されています。検索データベースは定期的に更新され、  
最新の情報が記録されます。IntelligentSearch は検索データベースにより、高速検索を  
実現しています。

【図11】

## 本発明の検索に使用するHTMLファイルに設けたプロパティ情報の説明図

プロパティ	
全般	
文書管理システムについて	
プロトコル Hyper Text 転送プロトコル (HTTP)	
種類	ファイル
アドレス (URL)	http://www.fujitsu.co.jp/productinst/view/document_001_buncyokanri
サイズ	3200 バイト
作成日	2000年5月16日
変更日	2000年5月16日
[証明] [分析]	
[OK] [キャンセル]	

【図12】

図8の処理によりテキスト抽出対象とするExcel文書の説明図

Microsoft Excel -Address.xls												
<u>ファイル(F)</u> <u>編集(E)</u> <u>表示(V)</u> <u>挿入(I)</u> <u>書式(O)</u> <u>ツール(T)</u> <u>データ(D)</u> <u>ウンドウ(W)</u> <u>ヘルプ(H)</u>												
A	B	C	D	E	F	G	H	I	J	K	L	M
1 氏名	連名	会社	部署	役職	郵便番号	住所1	住所2	住所3	電話番号	FAX番号	電子メールアドレス	備考
2												
3												
4												
5												
6												
7												
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
△ △ △ △ Sheet1 Sheet2 Sheet3												
コマンド												

【図13】

図12のExcel文書から抽出したテキスト文書の説明図

抽出テキスト文書

"Sheet" , "氏名" , "連名" , "会社" , "部署" , "役職" , "郵便番号" ,  
"住所\_1" , "住所\_2" , "住所\_3" , "電話番号" , "FAX番号" ,  
"電子メール\_アドレス" , "Sheet1" , "Sheet2" , "Sheet3"

【書類名】 要約書

【要約】

【課題】 検索データベースに登録されていない文書に類似した文書の検索を簡単な操作ですばやく行う。

【解決手段】 クライアント12からの検索要求に基づいてサーバ10で文書情報を検索して応答する装置であって、クライアント12の検索条件指定部26で検索条件に文書ファイルを指定した場合に、指定したファイル内容をネットワークを経由して送信する。サーバ10側に設けた検索マシン20の文書検索部30は、検索条件指定部26から送信されたファイル内容からキーワードを生成し、検索データベース22のインデックス（検索対象文書25から抽出した重要単語列）から類似文書を検索する

【選択図】

図1

出願人履歴情報

識別番号 [000005223]

1. 変更年月日 1996年 3月26日

[変更理由] 住所変更

住 所 神奈川県川崎市中原区上小田中4丁目1番1号  
氏 名 富士通株式会社